ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

A Survey on Facet Generation Framework Using Query Logs

Afreen Anjum T. khan, Prof. R. R. Keole

M.E Student, Department of Computer Science & Engineering, HVPM's College of Engineering & Technology, Amravati, India

Asst. Professor, Department of Information Technology & Engineering, HVPM's College of Engineering & Technology, Amravati, India

ABSTRACT: Web search queries are often ambiguous or multi-faceted, which makes a simple ranked list of results inadequate. To assist information finding for such faceted queries, we explore a technique that explicitly represents interesting facets of a query using groups of semantically related terms extracted from search results. As an example, for the query "baggage allowance", these groups might be different airlines, different flight types (domestic, international), or different travel classes (first, business, economy). We name these groups query facets and the terms in these groups facet terms. We develop a supervised approach based on a graphical model to recognize query facets from the noisy candidates found. The graphical model learns how likely a candidate term is to be a facet term as well as how likely two terms are to be grouped together in a query facet, and captures the dependencies between the two factors. We propose two algorithms for approximate inference on the graphical model since exact inference is intractable. Our evaluation combines recall and precision of the facet terms with the grouping quality. Experimental results on a sample of web queries show that the supervised method significantly outperforms existing approaches, which are mostly unsupervised, suggesting that query facet extraction can be effectively learned.

KEYWORDS*:* Query Facet, Faceted Search, Query Suggestion, Query Reformulation, Query Summarization.

I. INTRODUCTION

A query facet is a set of items which describe and summarize one important aspect of a query. Here a facet item is typically a word or a phrase. A query may have multiple facets that summarize the information about the query from different perspectives. For the query "watches", its query facets cover the knowledge about watches in five unique aspects, including brands, gender categories, supporting features, styles, and colors. The query "visit Beijing" has a facet about popular resorts in Beijing (tiananmen square, forbidden city, summer palace, ...) and a facet on several travel related topics (attractions, shopping, dining, ...).

Query facets provide interesting and useful knowledge about a query and thus can be used to improve search experiences in many ways. First, we can display query facet together with the original search results in an appropriate way. Thus, users can understand some important facets of a query without browsing tens of pages. For example, a user could learn different brands and categories of watches. We can also implement a faceted search
based on query facets. User can clarify their specific intent by selecting facet items. Then search results could be restricted to the documents that are relevant to the items. These multiple groups of query facets are in particular useful for vague or ambiguous queries, such as "apple". We could show the products of Apple Inc. in one facet and different types of the fruit apple in another. Second, query facets may provide direct information or instant answers that users are seeking. For example, for the query "lost season 5", all episode titles are shown in one facet and main actors are shown in another. In this case, displaying query facets can save browsing time. Third, query facets may also be used to improve the diversity of the ten blue links. We can re-rank search results to avoid showing the pages that are near-duplicated in query facets at the top. Query facets also contain structured knowledge covered by or related to the input keywords of a query, and thus they can be used in many other fields besides traditional web search, such as semantic

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

search or entity search. There has been a lot of recent work on automatically building knowledge ontology on the Web . Query facets can become a possible data source for this.

We observe that important pieces of information about a query are usually presented in list styles and repeated many times among top retrieved documents. Thus we propose aggregating frequent lists within the top search results to mine query facets and implement a system called QDMiner. More specifically, QDMiner extracts lists from free text, HTML tags, and repeat regions contained in the top search results, groups them into clusters based on the items they contain, then ranks the clusters and items based on how the lists and items appear in the top results. We propose two models, the Unique Website Model and the Context Similarity Model, to rank query facets. In the Unique Website Model, we assume that lists from the same website might contain duplicated information, whereas different websites are independent and each can contribute a separated vote for weighting facets. However, we find that sometimes two lists can be duplicated, even if they are from different websites.

II. LITERATURE REVIEW

E. Stoica, [1] Databases of text and text-annotated data constitute a significant fraction of the information available in electronic form. Searching and browsing are the typical ways that users locate items of interest in such databases. Faceted interfaces represent a new powerful paradigm that proved to be a successful complement to keyword searching. Automatic clustering techniques generate clusters that are typically labeled using a set of keywords, resulting in category titles such as “battery california technology mile state recharge impact official hour cost government”.

Ori Ben-Yitzhak, [2] Present a traditional faceted search to support richer information discovery tasks over more complex data models. Our first extension adds flexible, dynamic business intelligence aggregations to the faceted application, enabling users to gain insight into their data that is far richer than just knowing the quantities of documents belonging to each facet. The OLAP capabilities traditionally supported by database over relational data to the domain of free-text queries over metadata-rich content.

M. Diao, [3] To perform audio search in several languages, with very little resources being available in each language. The data was taken from audio content that was created in live settings and was submitted to the “spoken web” over a mobile connection. The “spoken web search” task of involves searching for audio content within audio content using an audio content query. The task required researchers to build a language- independent audio search system so that, given an audio query, it should be able to find the appropriate audio file(s) and the (approximate) location of query term within the audio file(s).

D. Dash, [4] A dynamic faceted search system for discovery- driven analysis on data with both textual content and structured attributes From a keyword query, we want to dynamically select a small set of “interesting” attributes and present aggregates on them to a user. Dynamic faceted search system for discovery driven analysis is often performed in On-Line Analytical Processing (OLAP) systems. From a potentially large search result, we want to automatically and dynamically discover a small set of facets and values that are deemed most “interesting” to a user. Dynamic faceted search system to support OLAP-style discovery driven analysis on a large set of structured and unstructured data.

M. J. Cafarella, [5] The World-Wide Web consists of a huge number of unstructured documents, but it also contains structured data in the form of HTML tables. We extracted 14.1 billion HTML tables from Google’s general-purpose web crawl, and used statistical classification techniques to find the estimated 154M that contain high-quality relational data. Because each relational table has its own “schema” of labeled and typed columns, each such table can be considered a small structured database. The effective techniques are used for searching for structured data at search-engine scales

T. Cheng, [6] as the Web has evolved into a data-rich repository, with the standard “page view,” current search engines is increasingly inadequate. While we often search for various data “entities” (e.g. phone number, paper PDF, date), today’s engines only take us indirectly to pages. Therefore, we propose the concept of entity search, a significant departure from traditional document retrieval. Towards our goal of supporting entity search, in the WISDM1 project at UIUC we build and evaluate our prototype search engine over a 2TB Web corpus. Our demonstration shows the feasibility and promise of large-scale system architecture to support entity search.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

H. Zhang, [7] a semantic class is a collection of items (words or phrases) which have semantically peer or sibling relationship. This paper studies the employment of topic models to automatically construct semantic classes, taking as the source data a collection of raw semantic classes (RASCs), which were extracted by applying predefined patterns to web pages. Appropriate preprocessing and postprocessing are performed to improve results quality, to reduce computation cost, and to tackle the fixed-k constraint of a typical topic model. An evaluation methodology for measuring the quality of semantic classes. We show by experiments that our topic modeling approach outperforms the item clustering and RASC clustering approaches.

Y. Hu, [8] Web search queries are often ambiguous or multi-faceted, which makes a simple ranked list of results inadequate. To assist information finding for such faceted queries, we explore a technique that explicitly represents interesting facets of a query using groups of semantically related terms extracted from search results. Search results clustering is a technique that tries to organize search results by grouping them into, usually labeled, clusters by query subtopics. Search results clustering is a technique that tries to organize search results by grouping them into, usually labeled, clusters by query subtopics. A supervised method based on a graphical model for query facet extraction. The graphical model learns how likely it is that a term should be selected and how likely it is that two terms should be grouped together in a query facet.

Weize Kong, [9] Peresent Faceted search enables users to navigate a multi-dimensional information space by combining keyword search with drill-down options in each facets. For example, when searching “computer monitor” in an e-commerce site, users can select brands and monitor types from the provided facets {“Samsung”, “Dell”, “Acer”,} and {“LET-Lit”, “LCD”, “OLED” ...}. It has been used successfully for many vertical applications, including e-commerce and digital libraries. We peresent both intrinsic evaluation, which evaluates facet generation on its own, and extrinsic evaluation, which evaluates an entire Faceted Web Search system by its utility in assisting search clarification. We also design a method for building reusable test collections for such evaluations. Our experiments show that using the Faceted Web Search interface can significantly improve the original ranking if allowed sufficient time for user feedback on facets.

L. Bing, [10] Peresent the model exploit salatent topic space, which is automatically derived from the query log, to detect semantic dependency of terms in a query and dependency among topics. We present a framework for performing query reformulation, and this framework mainly has the following contributions. First, the graphical model can detect and maintain consistency of semantic meaning when scoring a candidate query. It is also capable of conducting reliable scoring for the candidates of an unfamiliar query.

R. Baeza-Yates, [11] proposes a list of concerned queries. The concerned queries are founded in antecedently published queries, and can be published by the user to the search engine to tune or redirect the search process. The method proposed is based on a query clustering procedure in which groups of semantically like queries are named. The clustering procedure uses the content of historical preferences of users registered in the query log of the search engine. The method not only discloses the related queries, but also ranks them agreeing to a relevance criterion. As future work we tend to improve the notion of interest of the recommended queries and to develop alternative notions of interest for the question recommender system.

I. Szpektor, [12] in this paper, we focus on related query recommendation, one of the tasks for which the long-tail issue is the most visible. We propose to address the long-tail problem by leveraging query templates, which are query constructs that abstract and generalize queries. Our key idea is to identify rules between templates as means for suggesting related queries. In future work, we plan to apply the query-template flow graph on other search-related tasks and to further explore its structure and behavior. In addition, we want to improve the quality of rule extraction.

L. Li, L. Zhong, [13] present a Query-URL Bipartite based query recommendation approach, called QUBiC. It utilizes the connectivity of a query-URL bipartite graph to recommend related queries and can significantly improve the accuracy and effectiveness of personalized query recommendation systems comparing with the conventional pairwise similarity based approach.

Z. Zhang, [14] present a simple and intuitive method for mining search engine query logs to get fast query recommendations on a large scale industrial-strength search engine. In order to get a more comprehensive solution, we combine two methods together. On the one hand, we study and model search engine users’ sequential search behavior, and interpret this consecutive search behavior as client-side query refinement, that should form the basis for the search engine’s own query refinement process.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

Zhicheng Dou, [15] present Query reformulation and query recommendation (or query suggestion) are two popular ways to help users better describe their information need. Query reformulation is the process of modifying a query that can better match a user's information need, and query recommendation techniques generate alternative queries semantically similar to the original query. The main goal of mining facets is different from query recommendation. The former is to summarize the knowledge and information contained in the query, whereas the latter is to find a list of related or expanded queries. However, query facets include semantically related phrases or terms that can be used as query reformulations or query suggestions sometimes. Different from transitional query suggestions, we can utilize query facets to generate structured query suggestions, i.e., multiple groups of semantically related query suggestions. This potentially provides richer information than traditional query suggestions and might help users find a better query more easily. We will investigate the problem of generating query suggestions based on query facets in future work.

III.PROBLEM STATEMENT

The previous used methodologies were unable to get exact result because of incomplete information or insufficient data. Because of lot of facets, the system performance is very slow. Problem occurs in the association of facets i.e. maintaining relationship between the facets is a difficult task. The query subtopics are hidden from the user, leaving him or her to guess at how the results are organized.

- It may be difficult to formulate a good query if you don't know the query collection well.
- If user enters the long and noisy query into the system, the system returns incorrect data to the user.
- Mismatch of searcher's vocabulary versus collection vocabulary. If the user searches for laptop but all the documents use the term notebook computer, then the query will fail.
- The most critical language issue for retrieval effectiveness is the term mismatch problem: the indexers and the users do often not use the same word.
- Word inflection (such as with plural forms, "television" versus "televisions"), may result in a failure to retrieve relevant documents.

IV. MOTIVATION

Web search queries are often ambiguous or multi-faceted. Current popular approaches try to diversify the result list to account for different search intents or query subtopics. A weakness of this approach is that the query subtopics are hidden from the user, leaving him or her to guess at how the results are organized. In this work, we attempt to extract query facets from web search results to assist information finding for these queries. We define a query facet as a set of coordinate terms {i.e., terms that share a semantic relationship by being grouped to learns how likely a candidate term is to be a facet term as well as how likely two terms are to be grouped together in a query facet, and captures the dependencies between the two factors.

V. AIM & OBJECTIVES

Aim

To recognize query facets from the noisy facet candidate lists extracted from the top ranked search results.

Objectives

- The user issues a (short, simple) query.
- The system returns an initial set of retrieval results.
- The system computes a better presentation of the information need based on the user feedback.
- To learn interesting knowledge about the queries and multiple group of facets for better result.
- The system displays a best result on the top.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

VI.PROPOSED WORK

We are going to propose a systematic solution, which we refer to as, to automatically mine query facets by aggregating frequent lists from free text, HTML tags, and repeat regions within top search results. We create two human annotated data sets and apply existing metrics and two new combined metrics to evaluate the quality of query facets.

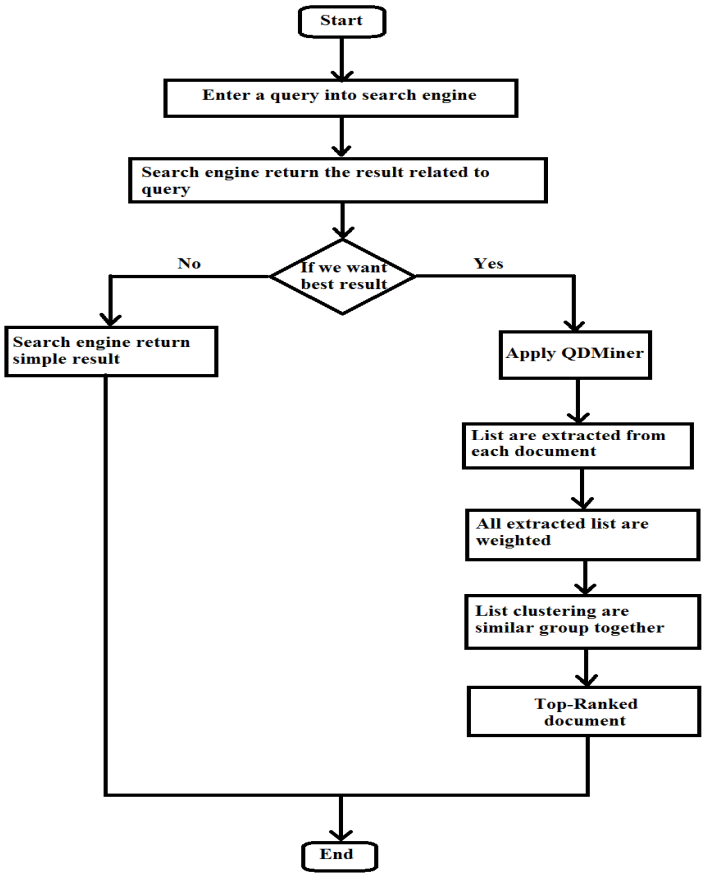

Fig.1 Flow Diagram

- **QDMiner**

QDMiner extracts lists from free text, HTML tags, and repeat regions contained in the top search results, groups them into clusters based on the items they contain, then ranks the clusters and items based on how the lists and items appear in the top results. The former is to summarize the knowledge and information contained in the query, whereas the latter is to find a list of related or expanded queries. QDMiner aims to offer the possibility of finding the main points of multiple documents and thus save users' time on reading whole documents. We implement a system called QDMiner which discovers query facets by aggregating frequent lists within the top results.

- **Working**

Step 1: List Extraction Several types of lists are extracted from each document in R. "men's watches, women's watches, luxury watches ..." is an example list extracted.

Step2: List Weighting All extracted lists are weighted, and thus some unimportant or noisy lists, such as the price list "299.99, 349.99, 423.99 ..." that occasionally occurs in a page, can be assigned by low weights.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

Step3: List Clustering Similar lists are grouped together to compose a dimension. For example, different lists about watch gender types are grouped because they share the same items "men's" and "women's".

Step4: Item Ranking Facets and their items are evaluated and ranked based on their importance. For example, the dimension on brands is ranked higher than the Facets on colors based on how frequent the dimensions occur and how relevant the supporting documents are. Within the Facets on gender categories, "men's" and "women's" are ranked higher than "unisex" and "kids" based on how frequent the items appear, and their order in the original lists.

VII. DESIRED IMPLECATION

We propose a systematic solution, which we refer to as QDMiner, to automatically mine query facets by aggregating frequent lists from free text, HTML tags, and repeat regions within top search results. We developed a supervised method based on a graphical model to recognize query facets from the noisy facet candidate lists extracted from the top ranked search results. We proposed two algorithms for approximate inference on the graphical model. We designed a new evaluation metric for this task to combine recall and precision of facet terms with grouping quality. Experimental results showed that the supervised method significantly out-performs other unsupervised methods, suggesting that query facet extraction can be effectively learned.

REFRENCES

[1] E. Stoica and M. A. Hearst, "*Nearly-automated metadata hierarchy creation*," in HLT-NAACL 2004: Short Papers, 2004, pp. 117–120.
[2] O. Ben-Yitzhak, N. Golbandi, N. Har'El, R. Lempel, A. Neumann, S. Ofek-Koifman, D. Sheinwald, E. Shekita, B. Sznajder, and S. Yogev, "*Beyond basic faceted search*," in Proc. Int. Conf. Web Search Data Mining, 2008, pp. 33–44.
[3] M. Diao, S. Mukherjea, N. Rajput, and K. Srivastava,, "*Faceted search and browsing of audio content on spoken web*," in Proc. 19th ACM Int. Conf. Inf. Knowl. Manage., 2010, pp. 1029–1038.
[4] D. Dash, J. Rao, N. Megiddo, A. Ailamaki, and G. Lohman, "*Dynamic faceted search for discovery-driven analysis,"* in ACM Int. Conf. Inf. Knowl. Manage., pp. 3–12, 2008.
[5] M. J. Cafarella, A. Halevy, D. Z. Wang, E. Wu, and Y. Zhang. "*Webtables: exploring the power of tables on the web,"* VLDB, 1:538–549, August 2008.
[6] T. Cheng, X. Yan, and K. C.-C. Chang. "*Supporting entity search: a large-scale prototype search engine,"* In Proceedings of SIGMOD '07, pages 1144–1146, 2007.
[7] H. Zhang, M. Zhu, S. Shi, and J.-R. Wen," *Employing topic models for pattern-based semantic class discovery,"* In Proceedings of ACL-IJCNLP '09, 2009.
[8] Y. Hu, Y. Qian, H. Li, D. Jiang, J. Pei, and Q. Zheng, "*Mining query subtopics from search log data"*, In Proceedings of SIGIR '12, pages 305–314, 2012.
[9] Weize Kong ,"*Extending Faceted Search to the Open-Domain Web"*, ACM SIGIR Forum Vol. 50 No. 1 June 2016.
[10] L. Bing, W. Lam, T.-L. Wong, and S. Jameel, "*Web query reformulation via joint modeling of latent topic dependency and term con- text,"* ACM Trans. Inf. Syst., vol. 33, no. 2, pp. 6:1–6:38, eb. 2015.
[11] R. Baeza-Yates, C. Hurtado, and M. Mendoza, "*Query recommendation using query logs in search engines,"* in Proc. Int. Conf. Cur- rent Trends Database Technol., 2004, pp. 588–596.
[12] I. Szpektor, A. Gionis, and Y. Maarek, "*Improving recommendation for long-tail queries via templates,"* in Proc. 20th Int. Conf. World Wide Web, 2011, pp. 47–56.
[13] L. Li, L. Zhong, Z. Yang, and M. Kitsuregawa, "*Qubic: An adaptive approach to query-based recommendation,"* J. Intell. Inf. Syst., vol. 40, no. 3, pp. 555–587, Jun. 2013.
[14] Z. Zhang and O. Nasraoui, "*Mining search engine query logs for query recommendation,"* in Proc. 15th Int. Conf. World Wide Web, 2006, pp. 1039–1040.
[15] Zhicheng Dou, Zhengbao Jiang, Sha Hu, Ji-Rong Wen, and Ruihua Song," *Automatically Mining Facets for Queries from Their Search Results", IEEE Transactions On Knowledge And Data Engineering*, Vol. 28, No. 2, February 2016.

BIOGRAPHY

Afreen Anjum is a student of M.E Computer Science & Engineering Department, H.V.P.M'S College of Engineering & Technology, Amravati, Maharashtra. She received Bachelor of Engineering Degree in 2015 from SGBAU Amravati, Maharashtra, India. Her research interests are Education technology and Data Mining.

Prof. R. R. Keole is a Asst. Professor in Department of Information Technology & Engineering, H.V.P.M'S College of Engineering & Technology, Amravati, Maharashtra, India.